Quantum and Quantum-Inspired Computation for NextG Wireless Baseband Processing

Kyle Jamieson

PAWS
Princeton Advanced Wireless Systems Lab

PRINCETON UNIVERSITY

With collaborators: John Kaewell (Interdigital), Srikar Kasi (Princeton), Abhishek Kumar (Princeton), Minsung Kim (Princeton), Aaron Lott (USRA), Salvatore Mandra (NASA Ames), Peter McMahon (Cornell), Davide Venturelli (USRA), Paul Warburton (Univ. College London)

NSF *Quantum-Enabled Networks* (QENeTs) Project (CNS-1824357, CNS-1824470)

NextG Evolution: Technologies Push, Demands Pull

Massive and Large MIMO

Error Control Codes

Wireless SDN

NOMA

Millimeter Wave

Ultra-Densification

Full-Duplex

Sub-6 GHz bands

5G/6G Wireless Technologies

Extreme Capacity:
> 10 Tbps

Massive IoT:
>1,000 nodes/ENodeB

Extreme Mobility:
≤ 500km/h

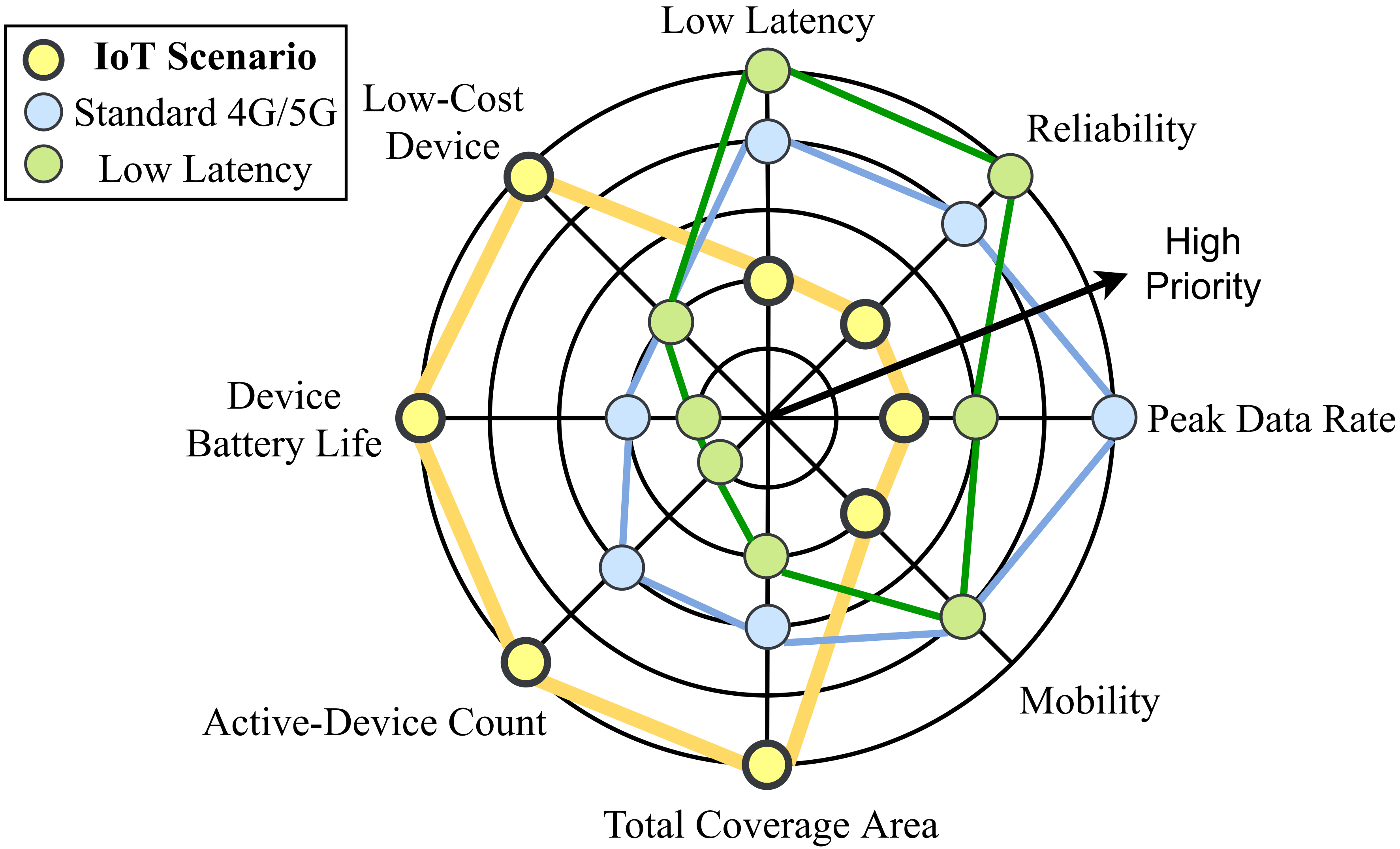
5G/6G Performance Goals

Data Rates:
> 100 Mbps/user

Ultra-Low Latency:
< 1 ms

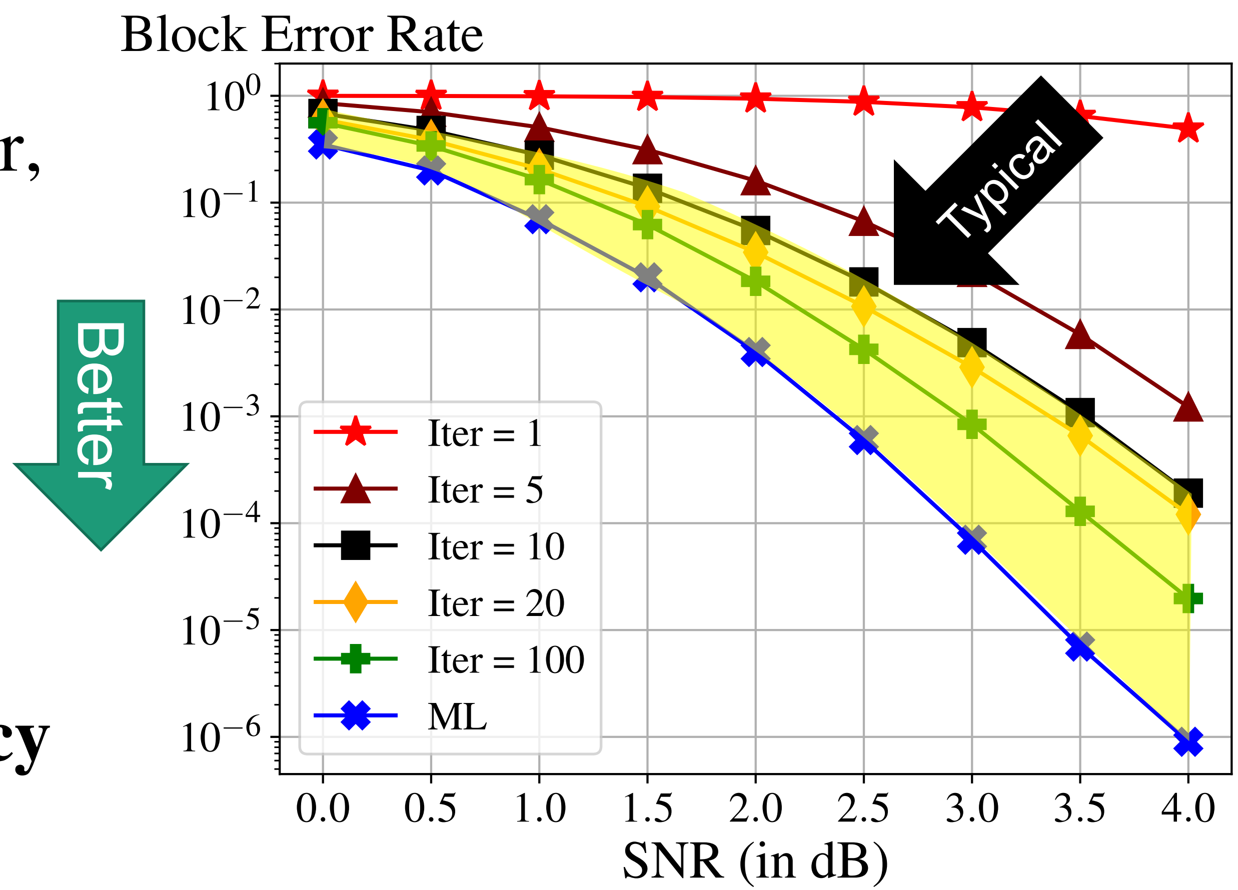
Ultra-low Energy:
>10x battery, green base stations

2G 3G 4G 5G+beyond



Status Quo Leaves Performance on the Table (1)

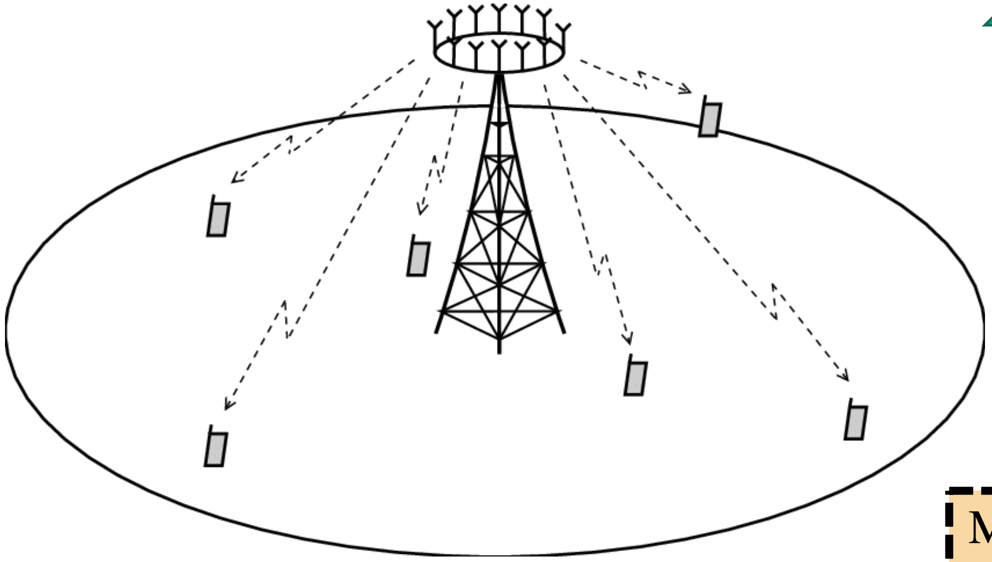
- **Example:** Belief Propagation decoder, $(155, 60)$ LDPC code
- Typical: 8-10 iterations
- **Two orders lower** block error rate is possible → **Higher spectral efficiency** (network capacity)

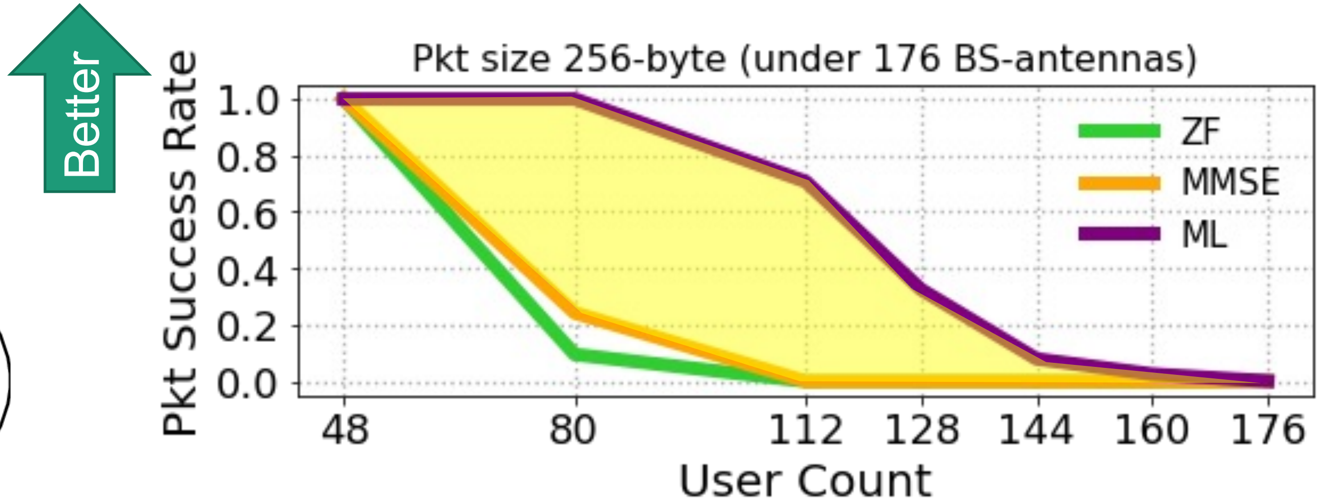


ML = Maximum Likelihood (best possible performance)

Status Quo Leaves Performance on the Table (2)

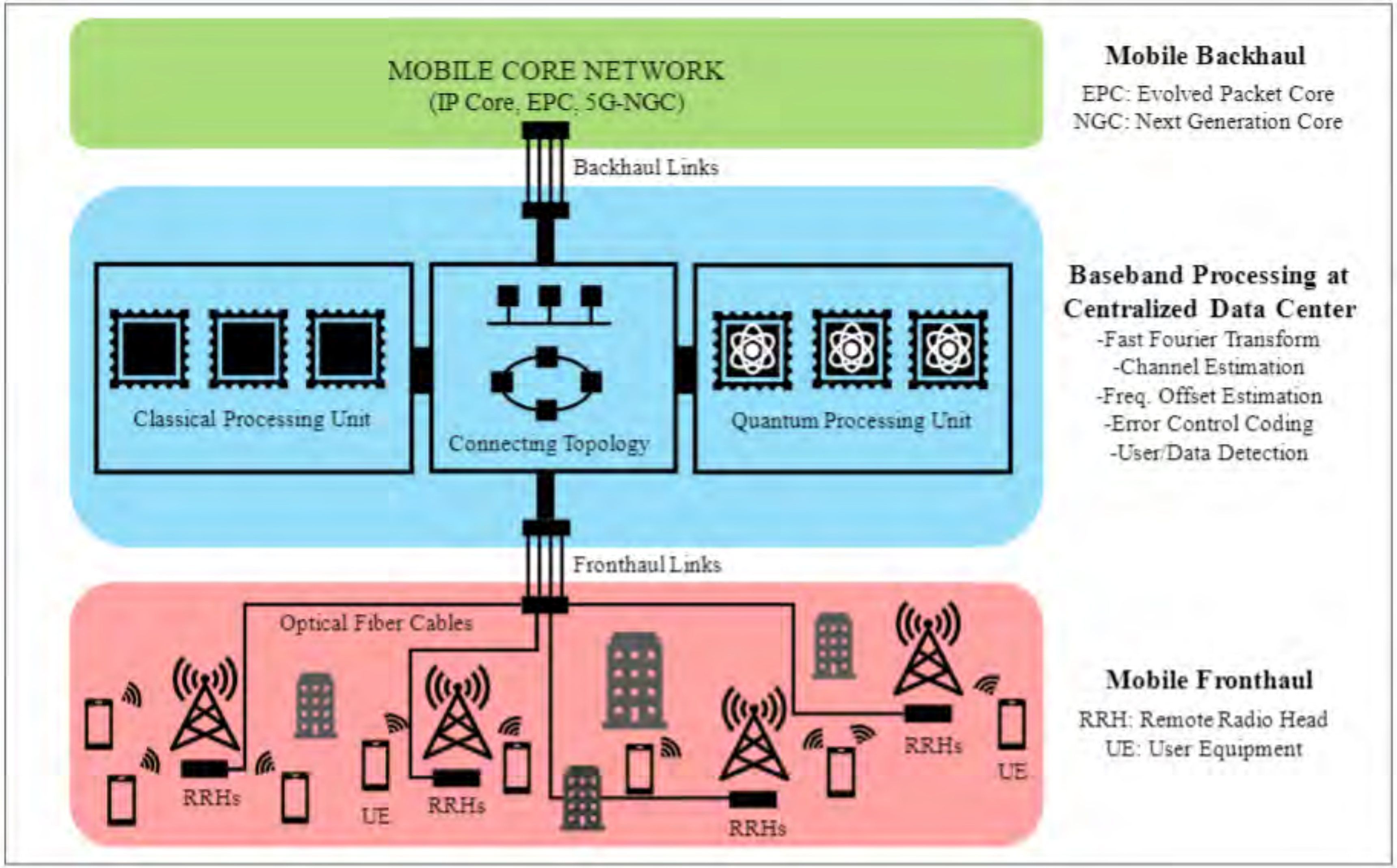
- **Example:** Multi-User Massive MIMO Detection
- Typical: Minimum Mean-Squared Error Receiver (MMSE)
- **Many-fold throughput gains** possible for 80-100 users (176-antenna base station)





ML = Maximum Likelihood (best possible performance)

Vision: Bring Quantum Processing to Baseband Units

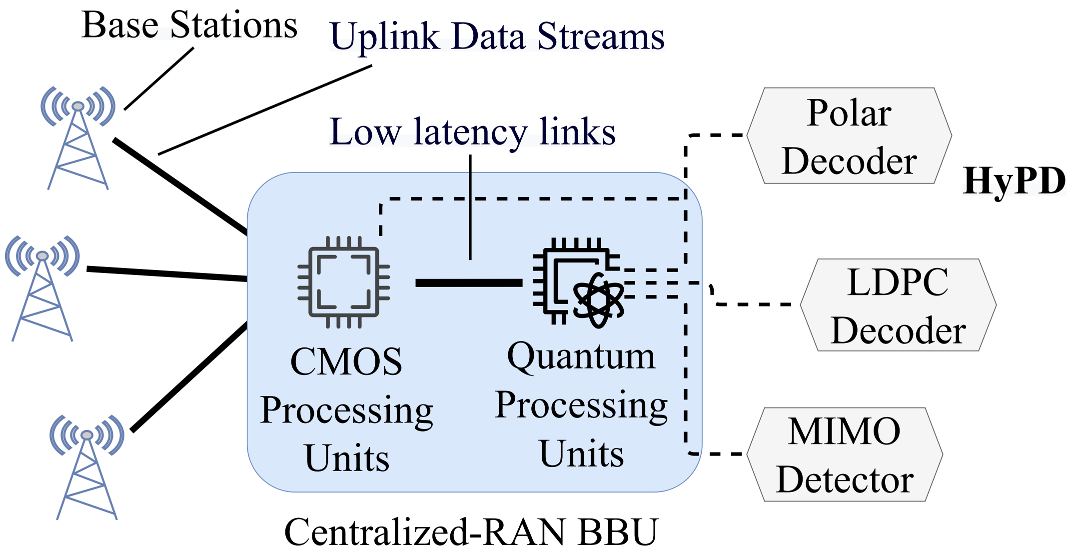


Quantum-Enabled Wireless Networks

- **Identify and evaluate the bottlenecks** to wireless capacity improvements
  - Algorithms
  - Hardware

- Investigate Quantum computation
  - Quantum Annealing
  - Quantum-Classical Hybrid
  - Quantum Gate model

- Make head-to-head performance comparisons
  - System cost, spectral efficiency, energy efficiency



Outline

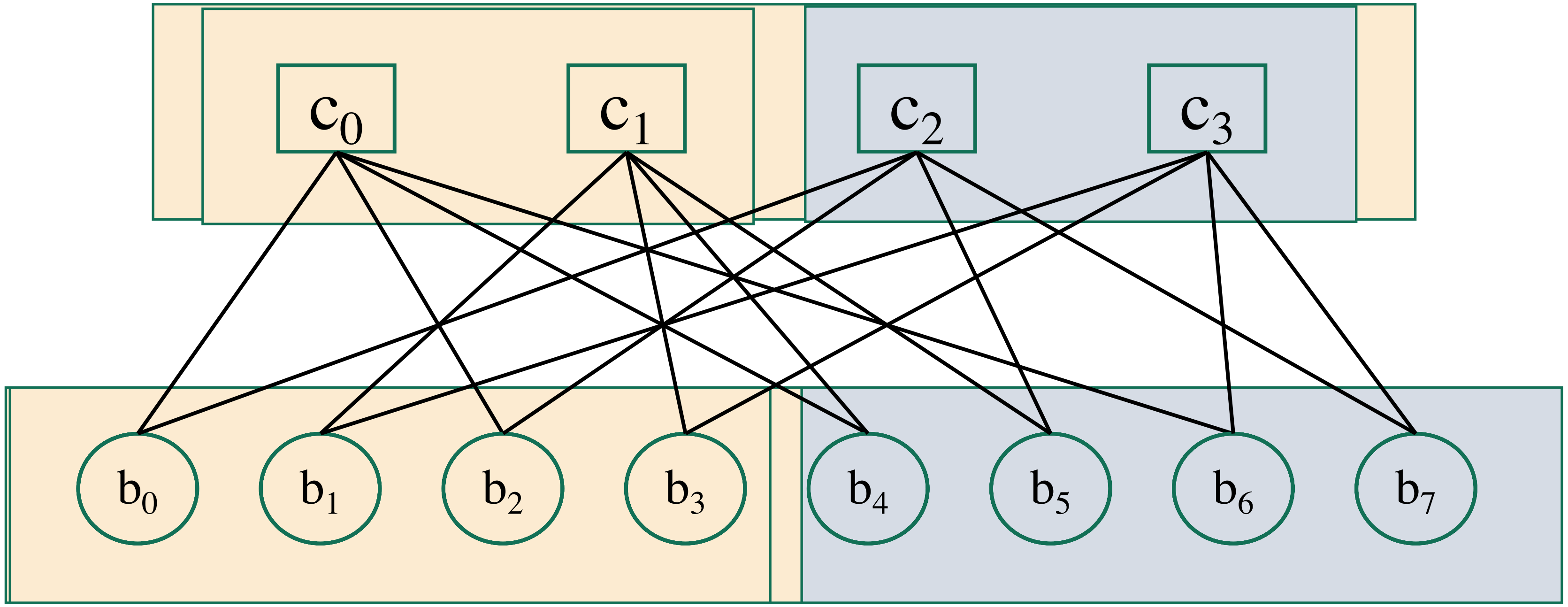
1. **Quantum LDPC decoder (*QBP*, MobiCom'20)**

2. Energy-performance analysis (ISCA QRE, arXiv '22)

3. Uplink MU-MIMO detection via Reverse Annealing (*IoT-ResQ*, MobiCom '22)

LDPC Decoding Status Quo: *Belief Propagation*

- Hardware (FPGAs/ASICs): Decoding Parallelism



- Fully parallel decoder
- Partially-parallel decoder
- Fully sequential decoder

Limitations of classical LDPC decoding

- Decoded via the *belief propagation (BP)* algorithm on FPGA/ASIC hardware
    - Accurate decoding = **high likelihood bit precision** (more resources)
    - Greater throughput = **high decoding parallelism** (more resources)
    - BP algorithm requires several **serial iterations** (impedes throughput)
- Network designers compromise between decoder accuracy and throughput
    - Fully parallel decoders with 8-bit precision (xcvu440 FPGA)
    - A (2,3)-regular code, block length 1944 bits, covers 72% of resources
    - A (4,8)-regular code, block length 2048 bits, exhausts resources

Primer: Quantum Annealing

- Analog/continuous interactions between superconducting qubits

- **Input:** Quadratic Unconstrained Binary Optimization (QUBO) problem

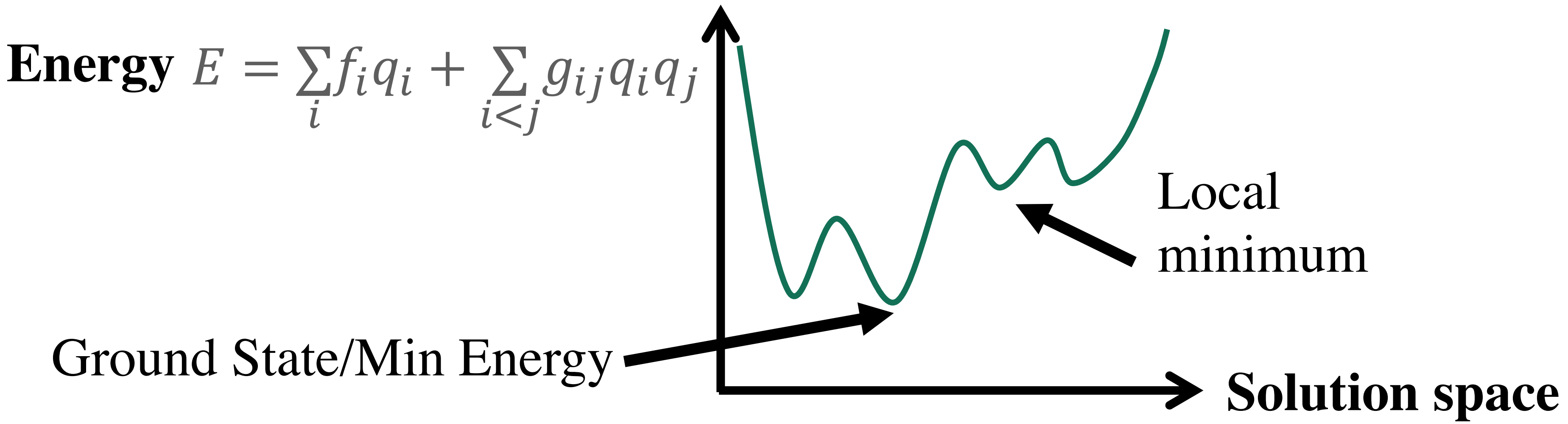
$$\hat{q}_1, \ldots, \hat{q}_N = \arg \min_{\{q_1, \ldots, q_N\}} \sum_{i \leq j}^{N} Q_{ij} q_i q_j$$

Programmed into QA hardware

- **Output:** Minimum energy **solution** of the QUBO problem $\widehat{q_1}, \ldots, \widehat{q_N}$

- Example: $\mathrm{Q} = \begin{bmatrix} Q_{11} & Q_{12} \\ 0 & Q_{22} \end{bmatrix} = \begin{bmatrix} 2 & -4.5 \\ 0 & 0.5 \end{bmatrix}$ → $2q_1 + 0.5q_2 - 4.5q_1q_2$

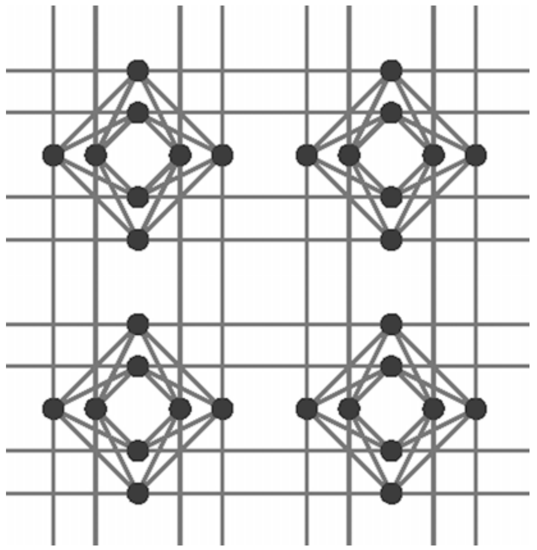
Quantum Annealing: Machine Runs



- **Anneal:** Single execution, tries to find the ground state or min energy solution
- **Anneal Time:** Duration of one anneal
- Need multiple anneals (one ***run***) to avoid local minima → Number of Anneals
- **Total Compute Time** = (Number of Anneals) × (Anneal Time)

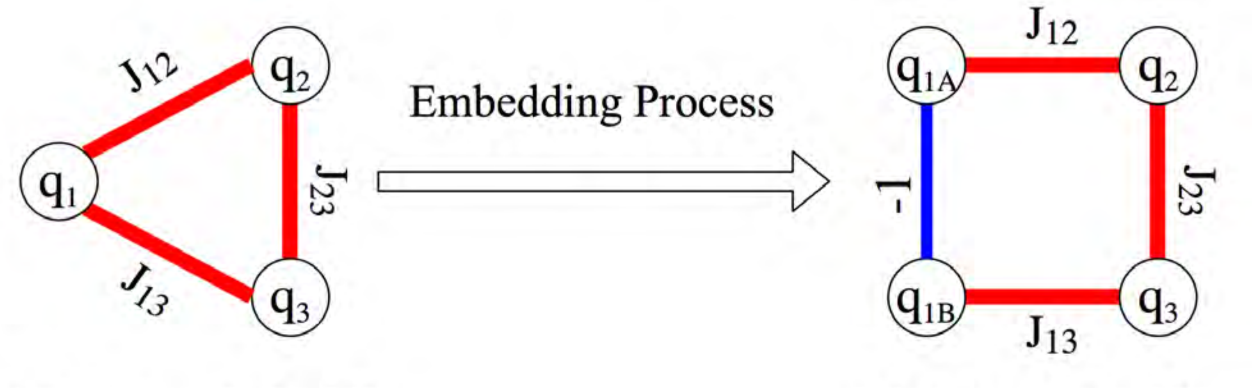
Quantum Annealing: Machine *Embedding*

QA hardware: Chimera Graph



Mapping a 3-variable fully connected problem

$$E = J_{12}\, q_1 q_2 + J_{13} q_1 q_3 + J_{23} q_2 q_3$$



(a) Before Embedding **(b)** After Embedding

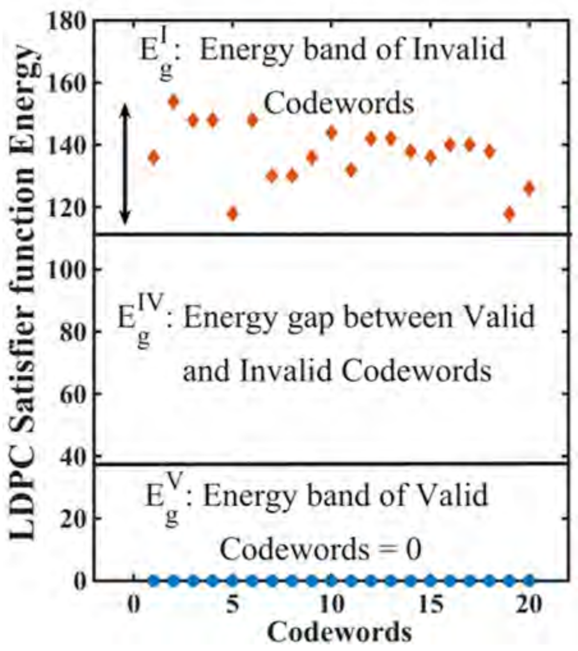
QA Workflow: Design a QUBO → **Map the QUBO onto QA hardware** → Solve the problem

Quantum Belief Propagation (QBP)

QUBO: $\min_{\mathbf{q}} \{ W_1 \sum_i (L_{sat}(c_i)) + W_2 \sum_j (\Delta_j) \}$

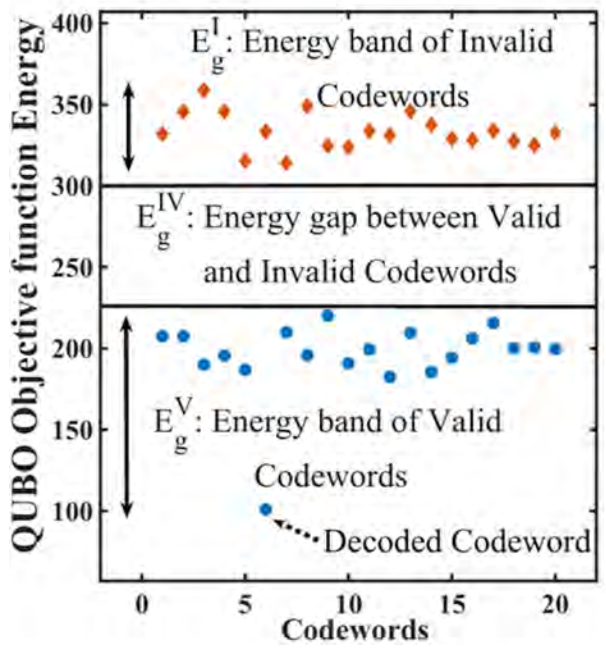
LDPC Satisfier

Ensures encoding



Distance

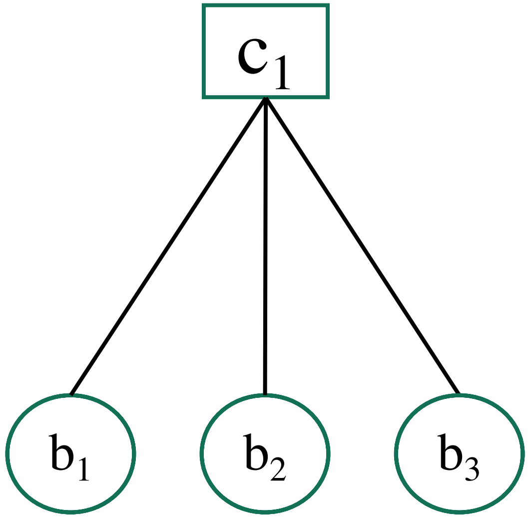
Finds correct answer



QBP: LDPC *Satisfier* function

- Encoding constraint : Modulo-two bit sum is zero at every check node

Example :



- c_1 checks three bits b_1, b_2, b_3
- Encoder Constraint: $b_1 \oplus b_2 \oplus b_3 = 0$ ➔ $b_1 + b_2 + b_3$ must be even
- Qubits for decoding $\{b_1, b_2, b_3\} = \{q_1, q_2, q_3\}$ respectively
- $L_{sat}(c_1) = (q_1 + q_2 + q_3 - 2q_{e1})^2$
- All q_i 's are binary variables. q_{e1} is ancillary.

QBP: LDPC *Distance* function

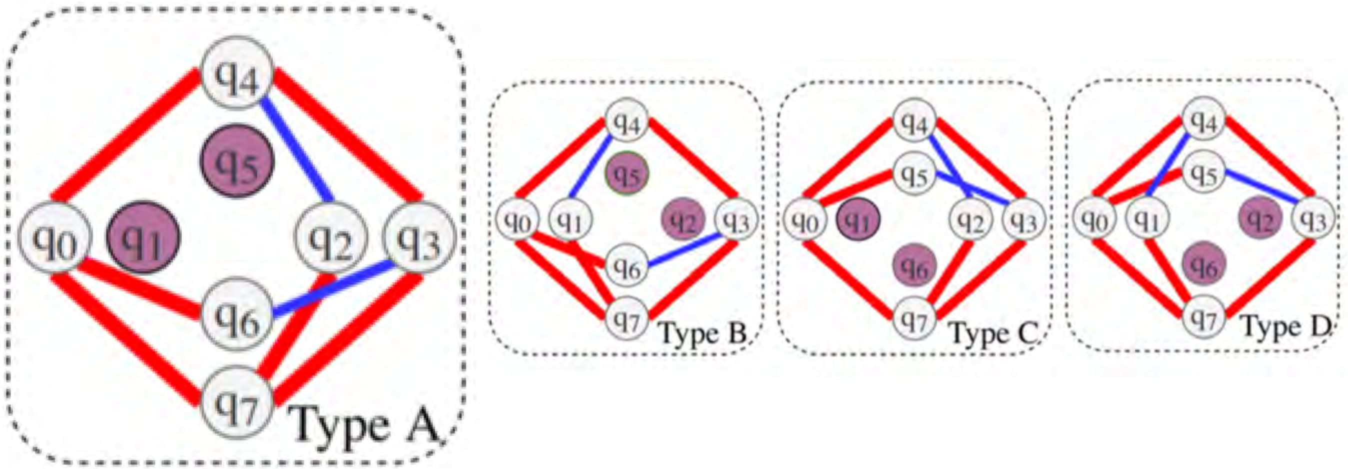
- Distance = proximity of candidate decoding to received information

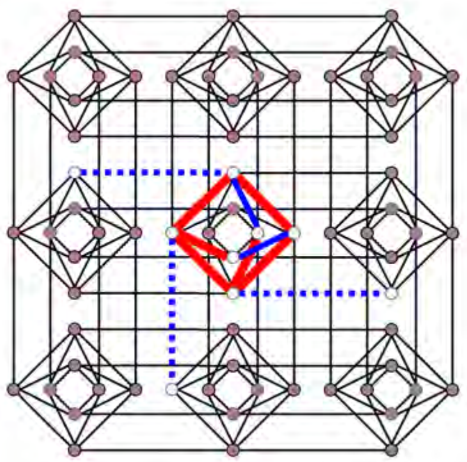
$$\Delta_i = (q_i - Pr(q_i = 1|y_i))^2$$

- qubit q_i corresponds to received bit y_i
- $\Delta_i \rightarrow$ minimal for a q_i in $\{0, 1\}$ $\rightarrow$ that has greater probability of being transmitted bit
- Probability is computed after soft demapping of received symbols

QBP's Embedding (Level-I)

- Two-Level Embedding.
- Example:
  - $L_{sat}(c_i) = (q_0 + q_4 + q_7 - 2q_{e3})^2$
- Construction:
  - Types A, B, C, D
- Placement:
  - One schema per unit cell
  - Shared bits placed closer

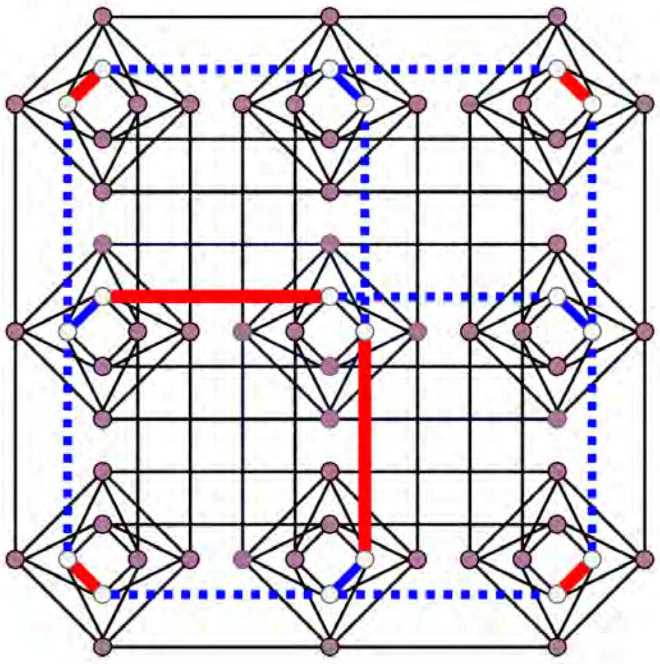




- Level-I embedding

QBP's Embedding (Level-II)

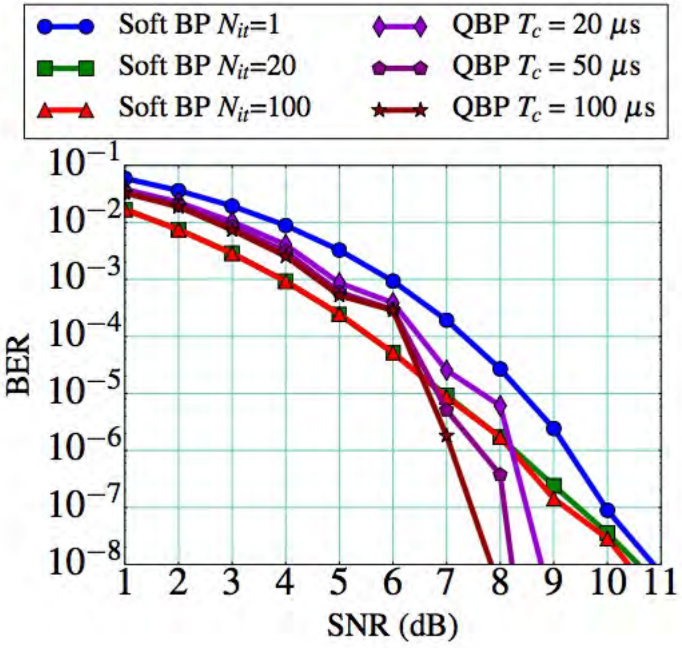
- Construction:
  - Based on Level-I placement
- Placement:
  - Shared bits placed closer
- QBP scales over entire hardware
- Every qubit is used efficiently.



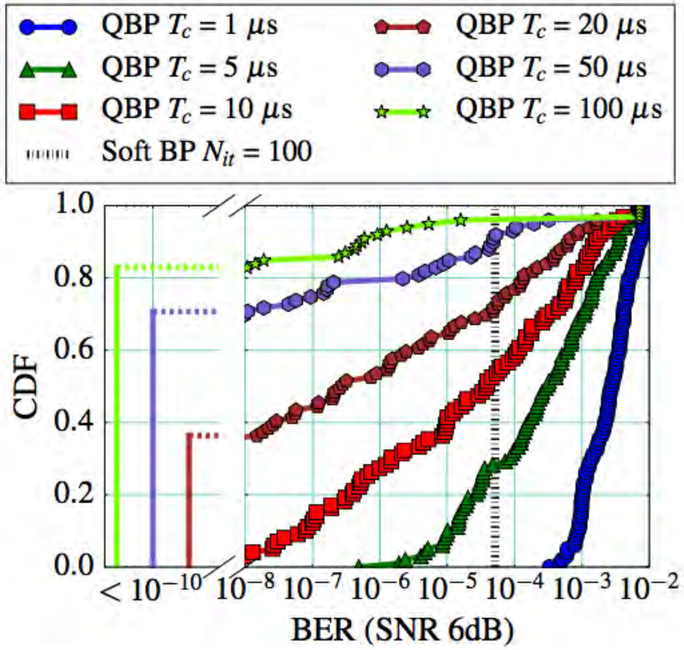
A	B	B
A	B	B
C	D	D

- Level-II embedding

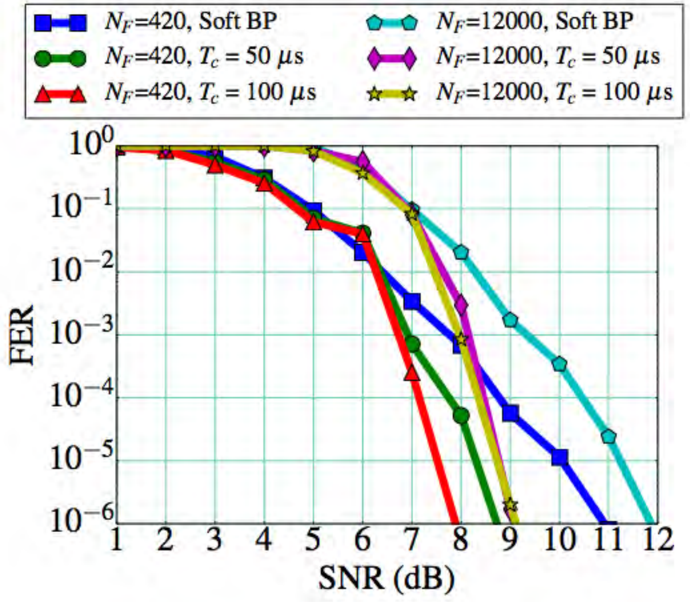
QBP: LDPC Decoding Error Performance



- Average BER



- Distribution of BERs



- Average FER

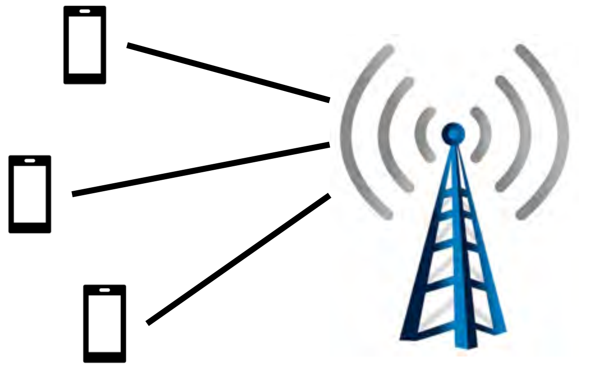
- QBP lags at SNRs < 6 dB, but reaches a 10^{-8} BER at 2-3 dB lower SNR than BP

A Cost and Power Feasibility Analysis of Quantum Annealing for NextG Cellular Wireless Networks

with Srikar Kasi, P. A. Warburton (University College London),
John Kaewell (InterDigital Corporation)

Motivation

Wireless Communication



User Equipment (UE)

Base Station (BS)

- **Internet users**
  - 2019: 3.9 billion users (51% of population)
  - 2023[1]: 5.3 billion users (66% of population)
- **Robust 5G technologies**
  - MIMO communication, Channel Coding
  - millimeter-Wave communication

Increasing Power Consumption

Economic (OpEx)

Environmental (Carbon emissions)

1. Cisco Annual Internet Report (2018-2023) White Paper

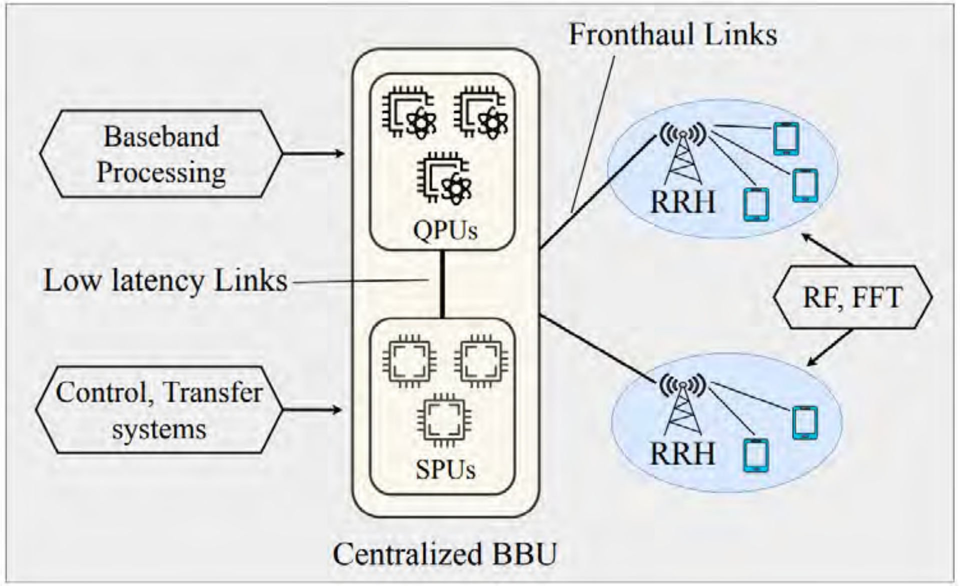
Controlling Power Consumption

- Sleep mode
  - Turn BS on/off during idle/low traffic times
- Optimize radio transmission
  - Approximate algorithms (Low complexity)
- **Improve hardware components**
  - CMOS hardware: ***Performance-per-Watt*** efficiency improving over years
  - But ***expected to terminate ca. 2030*** (End of Moore's Law)

Will CMOS achieve NextG cellular spectral and energy efficiency targets?

Envisioned Scenario

- **Centralized Radio Access Networks (C-RAN):**
  - Quantum Computation ⟶ Heavyweight tasks
  - Classical Computation ⟶ Lightweight tasks



- **Key Idea:**
  - Invest in ***Capital Expenditure*** (CapEx) ⬆
  - Reduce ***Operational Expenditure*** (OpEx) ⬇
  - Reduce ***Total Cost of Ownership (TCO)*** = CapEx + OpEx ⬇

Questions & Answers: Takeaways

Case Study: **Quantum Annealing (QA)** devices

1. **How many quantum bits (qubits) do we need for 5G processing?**
   a) Small BS → 40K qubits
   b) Macro BS → 3M qubits

✓ Highly Sparse Connectivity
✓ Multiple independent chips

2. **How much power/cost QA can save over CMOS?**
   a) Small BS → No benefit
   b) Macro BS → 41 kW (45% lower)

3. **At what year will these systems become feasible?**
   a) Small BS → *ca*. 2026 (best scenario)
   b) Macro BS → *ca*. 2036 (best scenario)

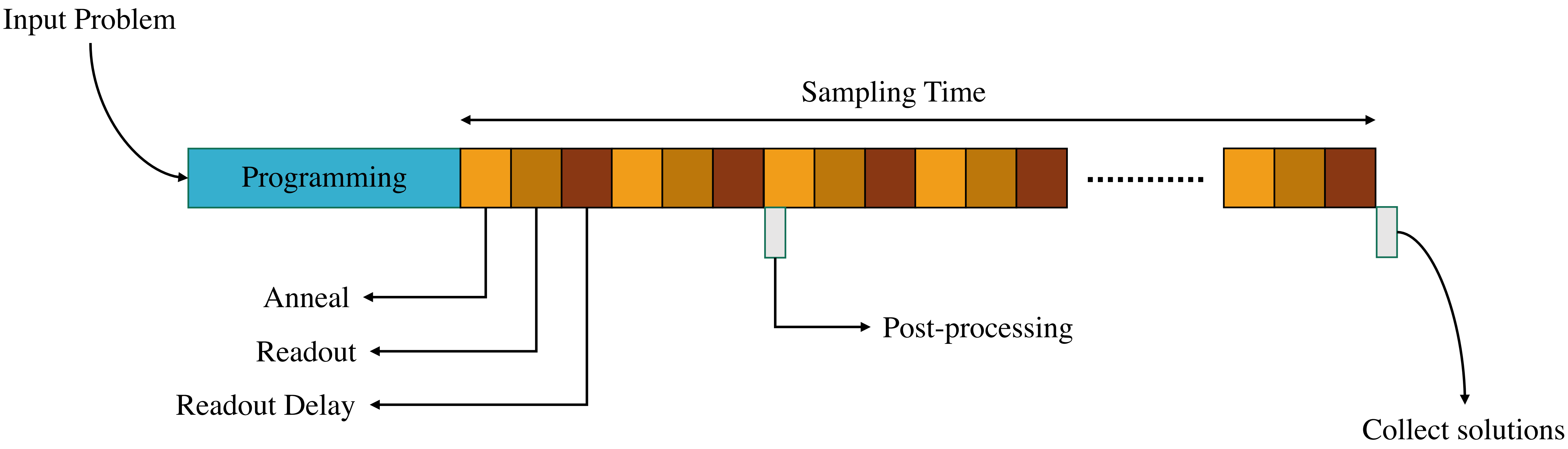
Evaluation: Methodology

- **Figures of Merit:**
  - Spectral Efficiency (bits/sec/Hz)
    - QA Latency
    - Number of qubits
  - Energy Efficiency (W/bit)
    - QA Power consumption
    - Number of qubits

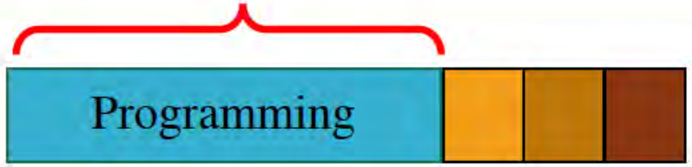
(Latency, Qubit count, and Power consumption) determine whether QA can benefit over CMOS

- **CMOS vs QA head-to-head, <u>at equal spectral efficiency</u>**

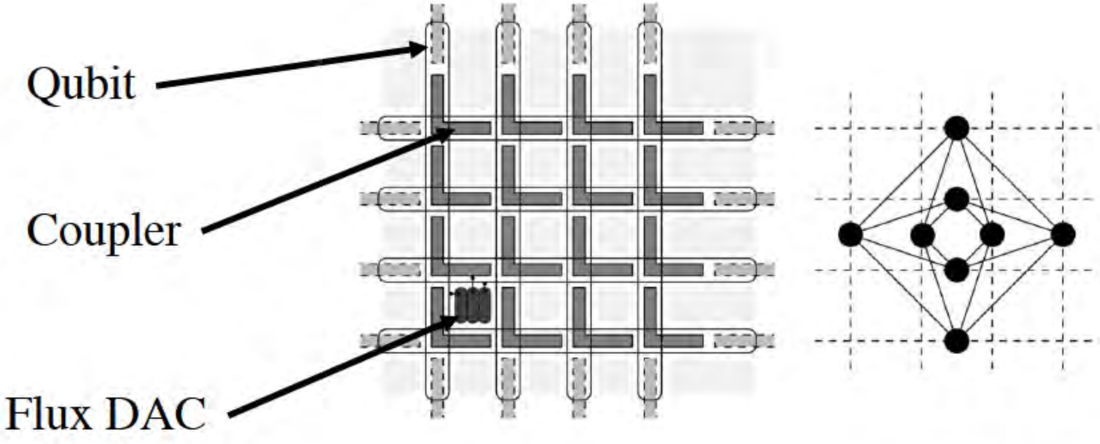
A Day in the Life of a QA Problem



QA Processing: Programming Phase



- Programming = Coefficients setting + Thermalization + Reset
  - Coefficients setting:
    - Time = 4 -- 40 μs
    - Bigger devices ⟶ More control line bandwidth
  - Thermalization:
    - 10M-qubit device: 36 pJ heat dissipation
    - QPU chip cooling power (15 mK) = 30 μW
    - Time = 1.2 μs
  - Reset:
    - Initialize qubits (Purcell Loss)
    - Qubit reset time[1] = 0.8 μs (99% confidence)
  - Overall Programming = 42 μs

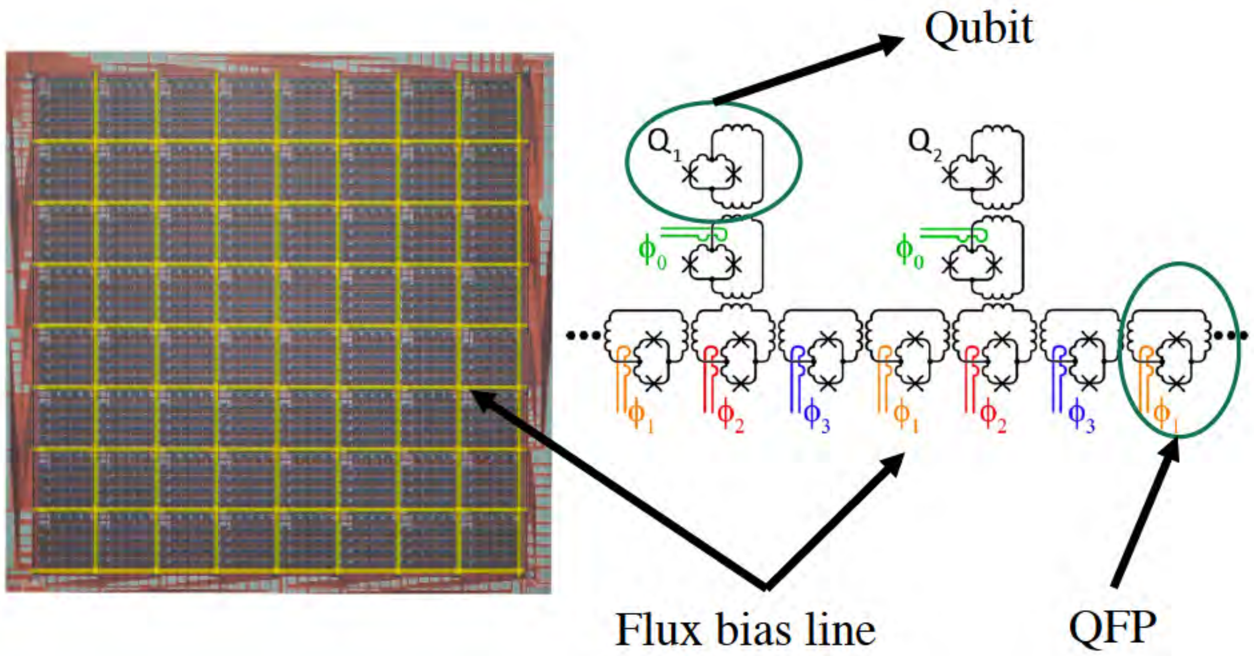


Physical Graphical

Bunyk *et al*, Architectural Considerations in the Design of a Superconducting Quantum Annealing Processor. TAS 2014.

QA Processing: Sampling Phase

- Anneal
  - Time = 1 μs
  - Dictated by control line bandwidth
- Readout
  - Time-division = 25 — 150 μs per sample
  - Frequency-multiplex = 1 μs per sample
- Readout Delay
  - Qubit Reset
  - Time = 1 μs per sample
- **For N_S samples:**
  - **Total Time = 42 + 3N_S μs**



Whittaker *et al,* A Frequency and sensitivity tunable microresonator array for high-speed quantum processor readout. Applied Physics 2016.

Estimating the Required Number of Qubits

Cellular Baseband Unit (BBU)

Frequency Domain

Forward Error Correction

Filtering

Equalization, etc.

89.6 TOPS[1] (Tera Operations per Second)

150M operations per problem (5G's Longest LDPC code)

600K PPS (Problems per Second)

21, 120 qubits per problem[2]

$42 + 3N_s\ \mu s$ per problem

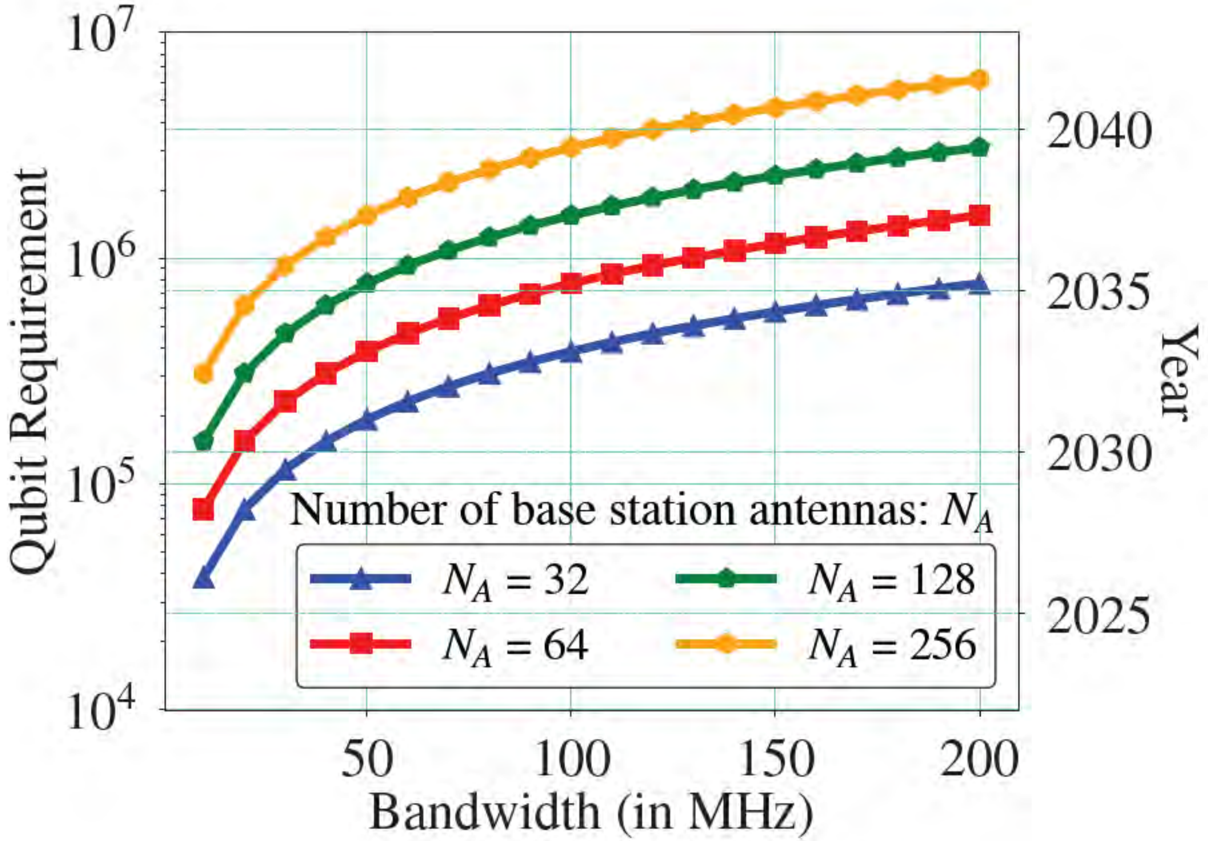
Example 5G Scenario:
200 MHz, 128 antennas,
64-QAM, 0.5 Code rate,
100% Time and
Frequency duty cycles

Qubit Requirement (FEC) = 600K/s * 21,120 * 102 μs = 1.3 M qubits

Claude Desset *et al*. Flexible Power Modeling of LTE base stations. IEEE Wireless Communications and Networking.

Srikar Kasi and Kyle Jamieson. Towards Quantum Belief Propagation for LDPC Decoding in Wireless Networks. ACM MobiCom.

Total Projected Qubit Requirement



Showing estimated year (at current growth trends) qubit requirement will be met

Power Consumption: Methodology

QA Hardware

- ➢ Power Consumption = 25 kW
  - Dominated by refrigeration unit
  - ***All qubits must fit in the same refrigeration unit***

- Tile of eight qubits (die) = 335 × 335 μm^2 QPU chip area
- QA experimental space = 250 mm radius

- Number of dies per wafer = 1.75M
- Number of qubits = 1.75M × 8 = 14M qubits

- ➢ 5G qubit count estimates are significantly lower
  - QA power consumption = 25 kW

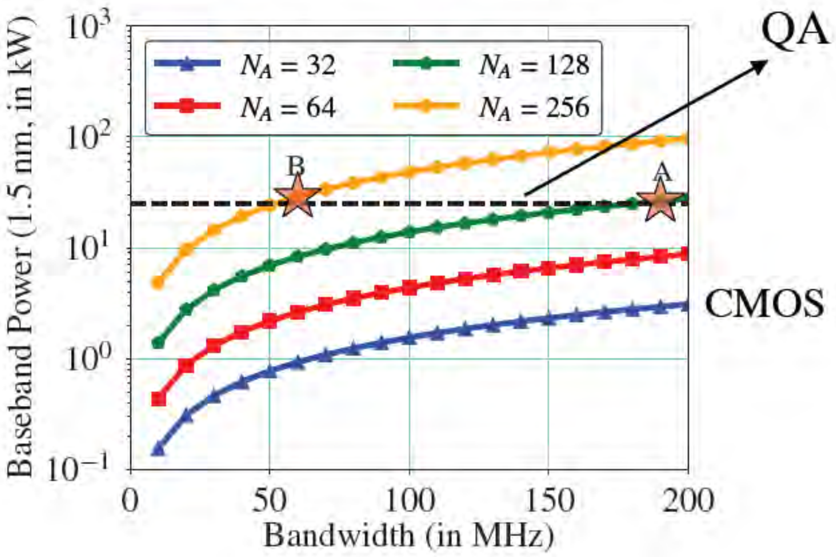
CMOS Hardware

- ➢ Power Consumption
  - Amount of computation (TOPS)
  - CMOS ***performance per Watt*** efficiency

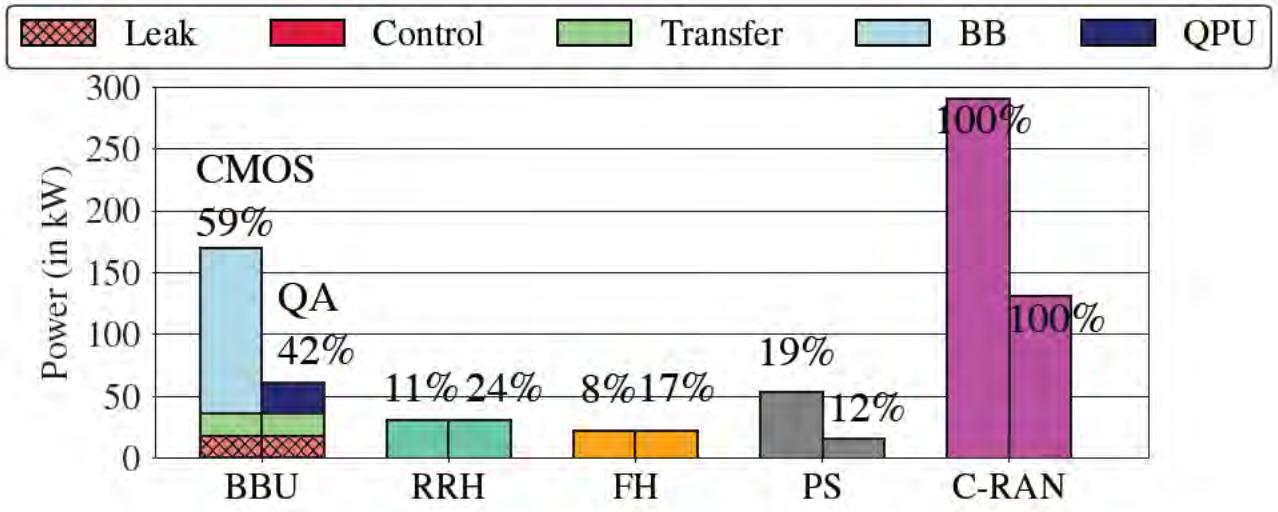
- Current 14nm CMOS = 0.076 TOPS/Watt
- Future 1.5nm CMOS (*ca*. 2030) = 0.3 TOPS/Watt

- ➢ Leakage Power = 30% of Dynamic Power

Power Comparison: QA vs CMOS



Point A: 190 MHz B/W, 128 antennas
Point B: 60 MHz B/W, 256 antennas

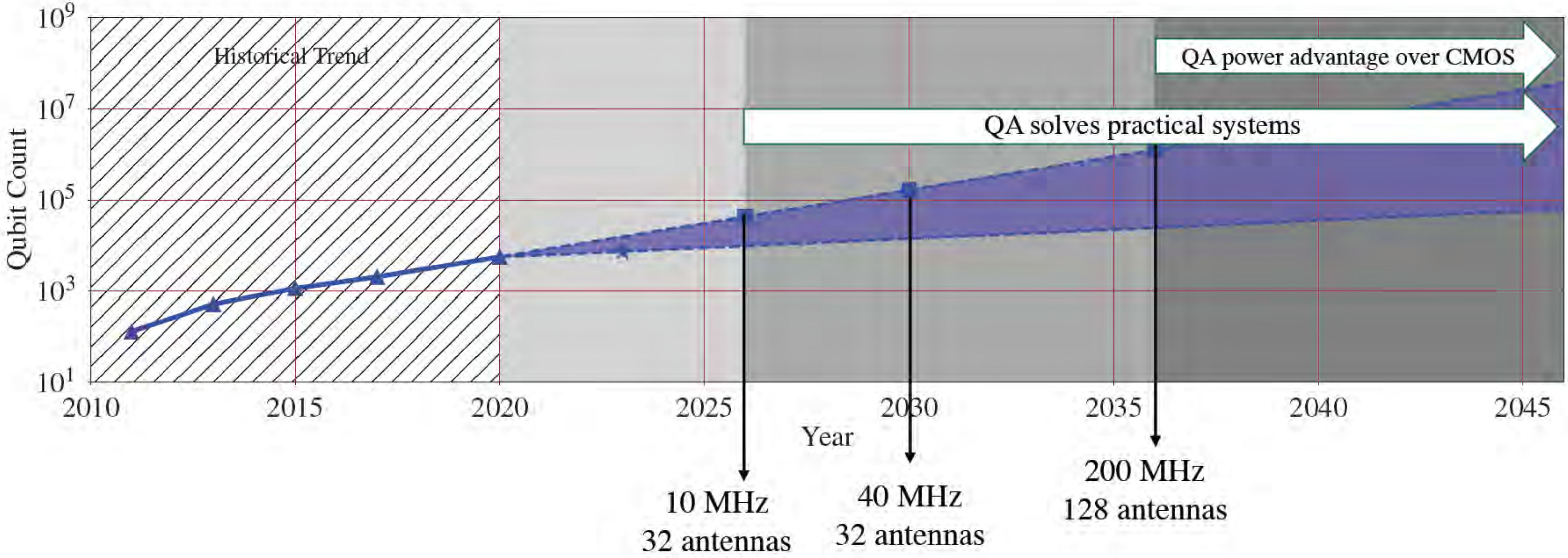


5G C-RAN: 3 BSs

Each BS: 400 MHz B/W, 128 antennas
- Non-linear MIMO
- mm-Wave communication

Left → CMOS
Right → QA

A Projected Feasibility Timeline



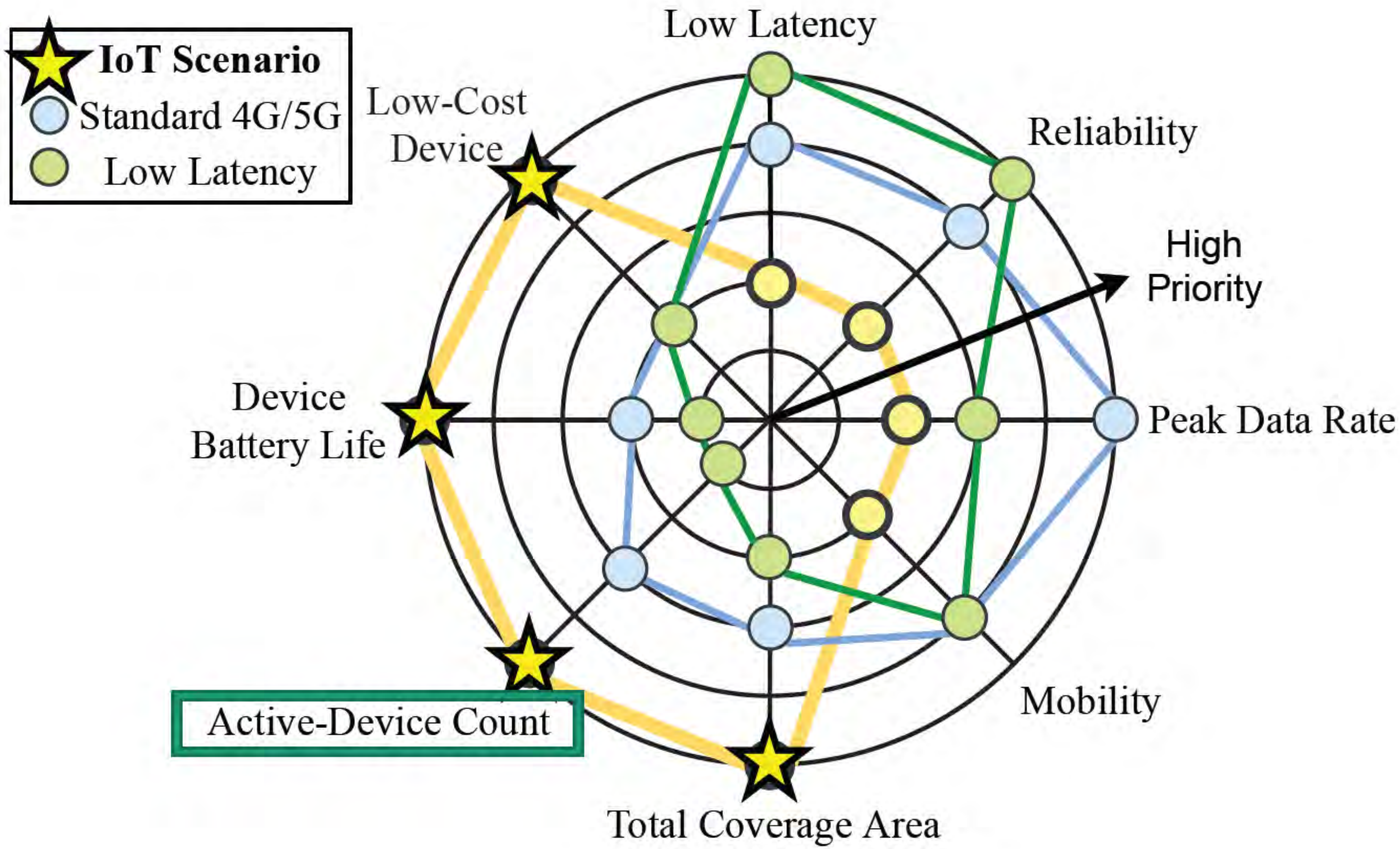
Outline

1. Quantum LDPC decoder (*QBP*, MobiCom’20)

2. Energy-performance analysis (ISCA QRE, arXiv ‘22)

3. **Uplink MU-MIMO detection via Reverse Annealing (*IoT-ResQ*, MobiCom ‘22)**

2G 3G 4G ⟫ IoT-ResQ: Goals



IoT-ResQ: High Connectivity Target

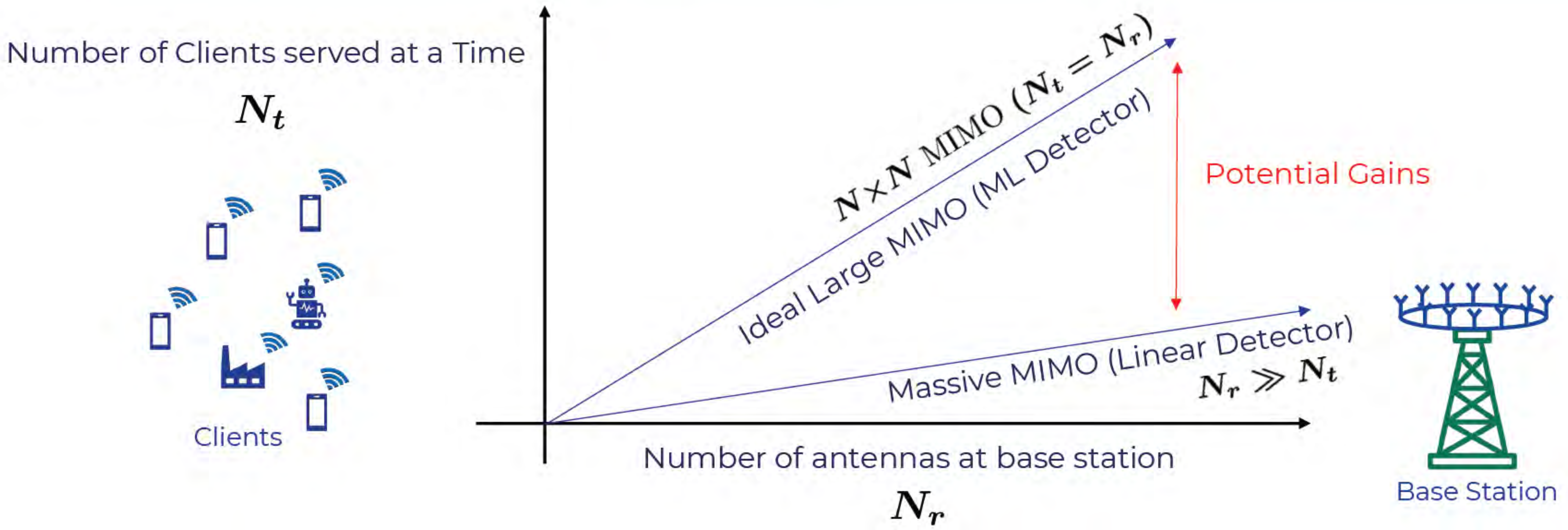
MU-MIMO Detection Methods

Linear Detection ex. Zero-Forcing (ZF)

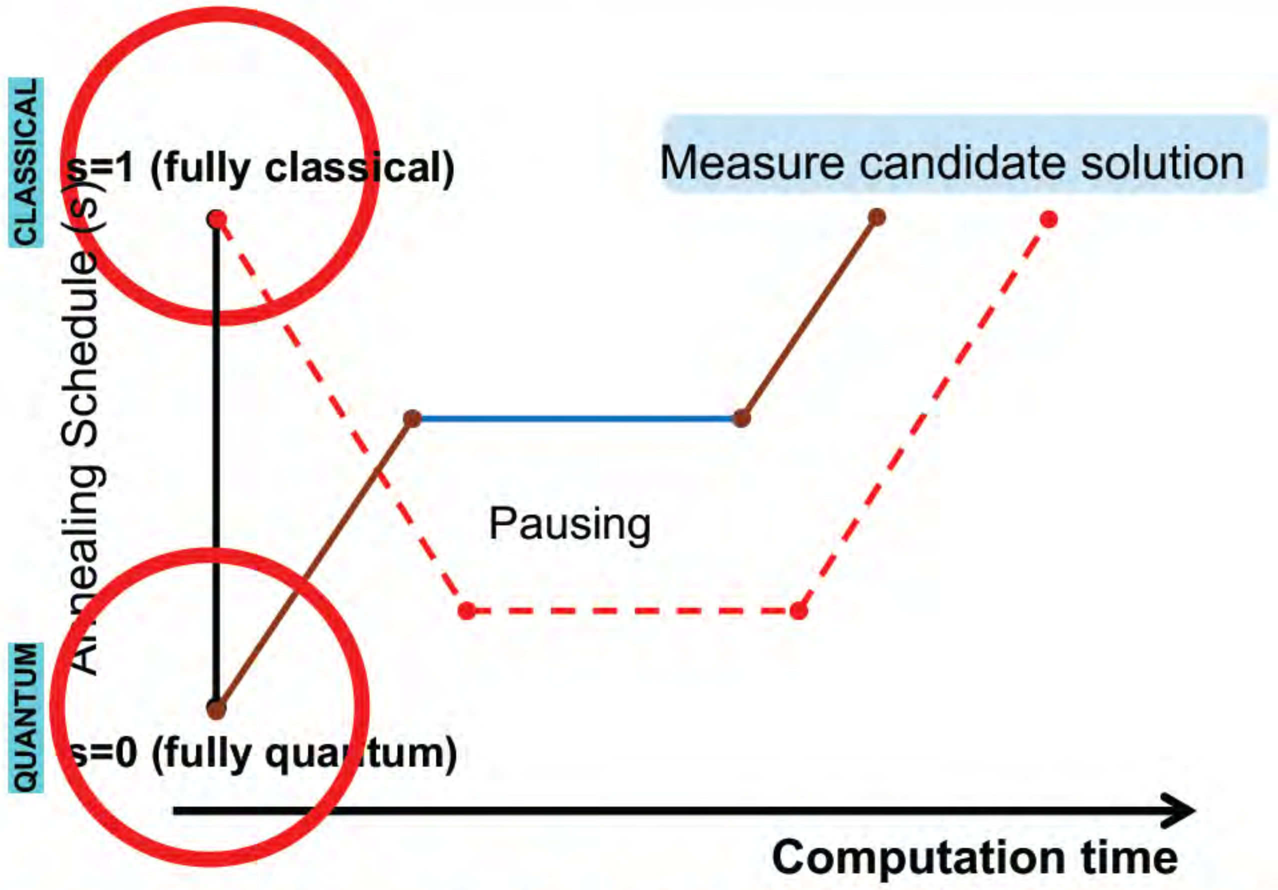
- Low Complexity and Easy Implementation
- Poorly perform for Large MIMO (poor channel conditions)

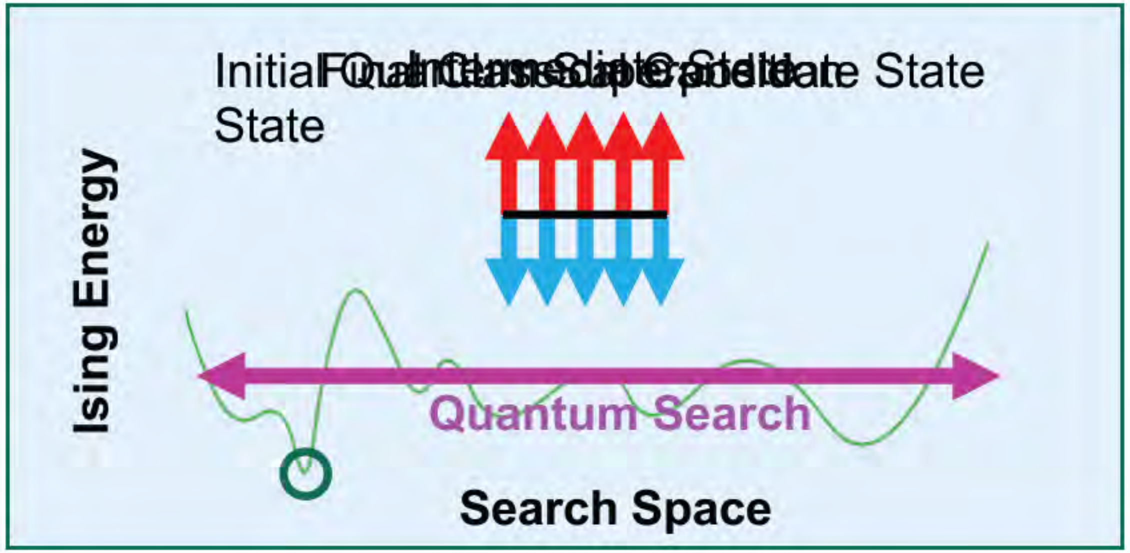
Maximum Likelihood (ML) Detection

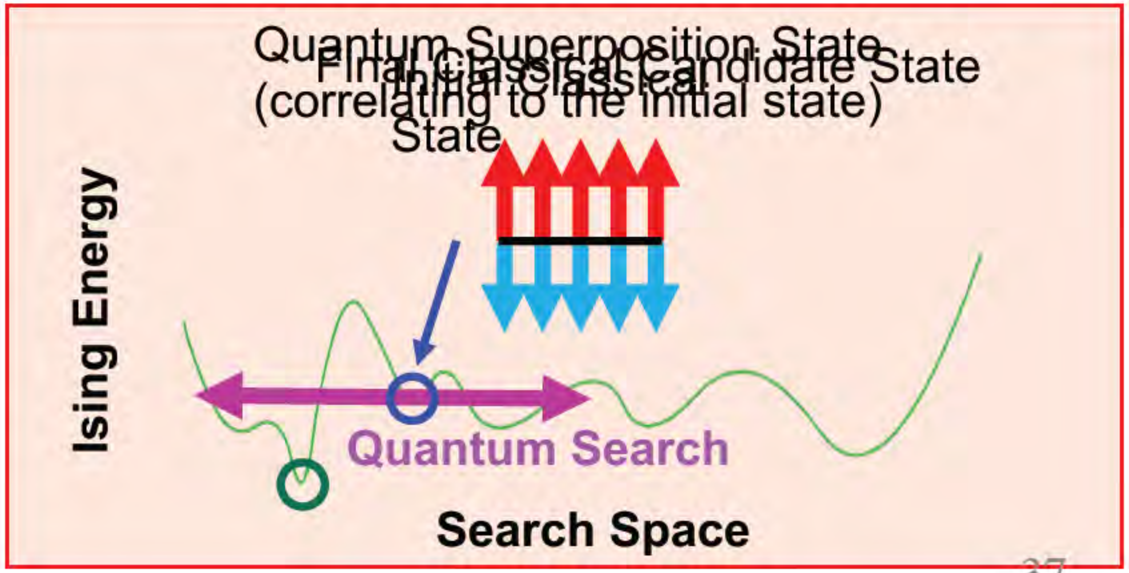
- Optimal MIMO Detection Performance (Lowest BER)
- Exponentially-Increasing Complexity



QA: Manipulating the Annealing Schedule



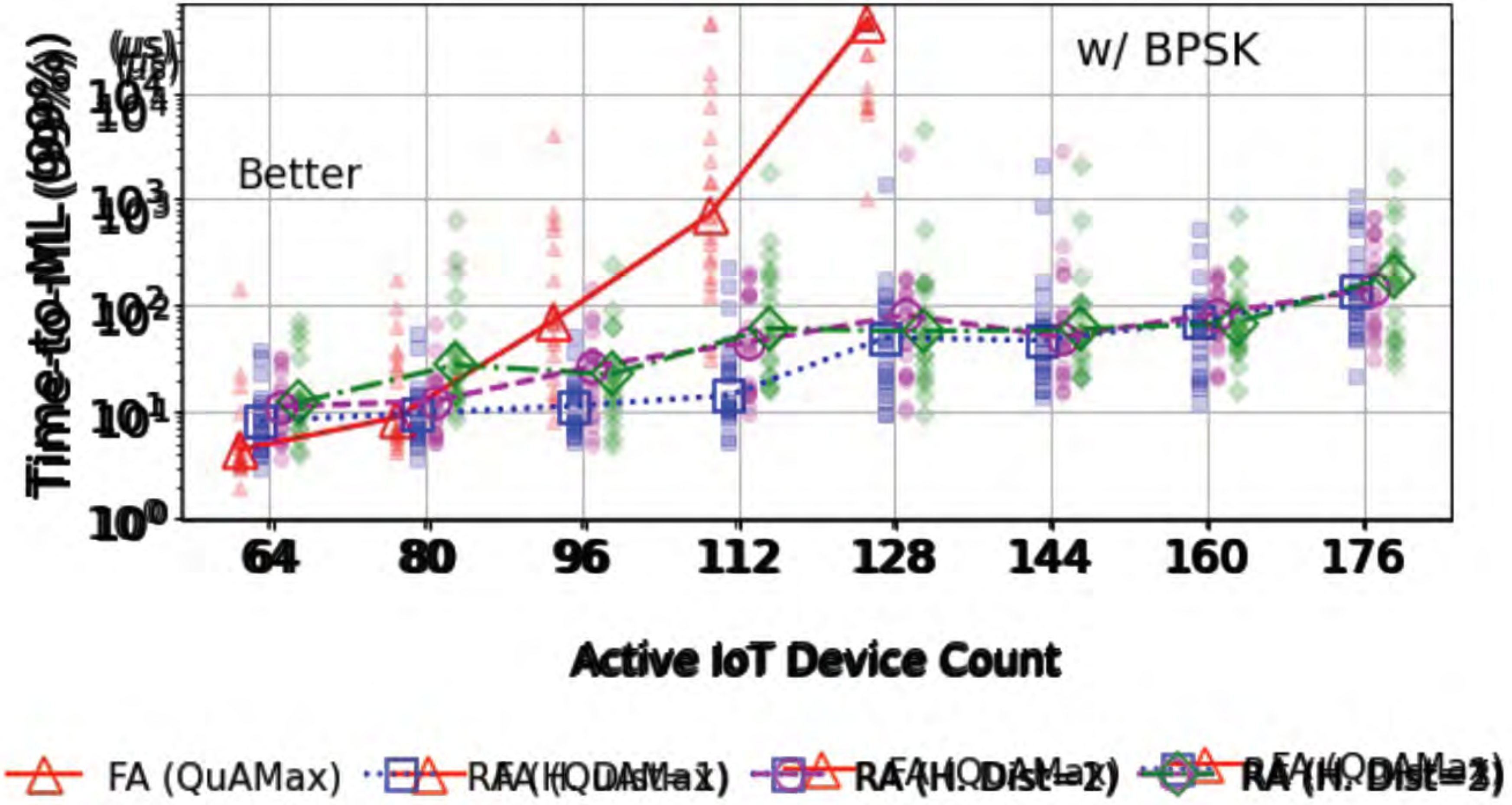




Forward Annealing (FA): QuAMax [SIGCOMM '19]

Reverse Annealing (RA): IoT-ResQ [MobiCom '22]

Performance Evaluation:
Forward Annealing (FA) vs Reverse Annealing (RA)



D-Wave Advantage

H. Dist = Inital State's Hamming Distance against the Optimal ML Sol.

Networking and Physics: Perspectives

The Networking Perspective

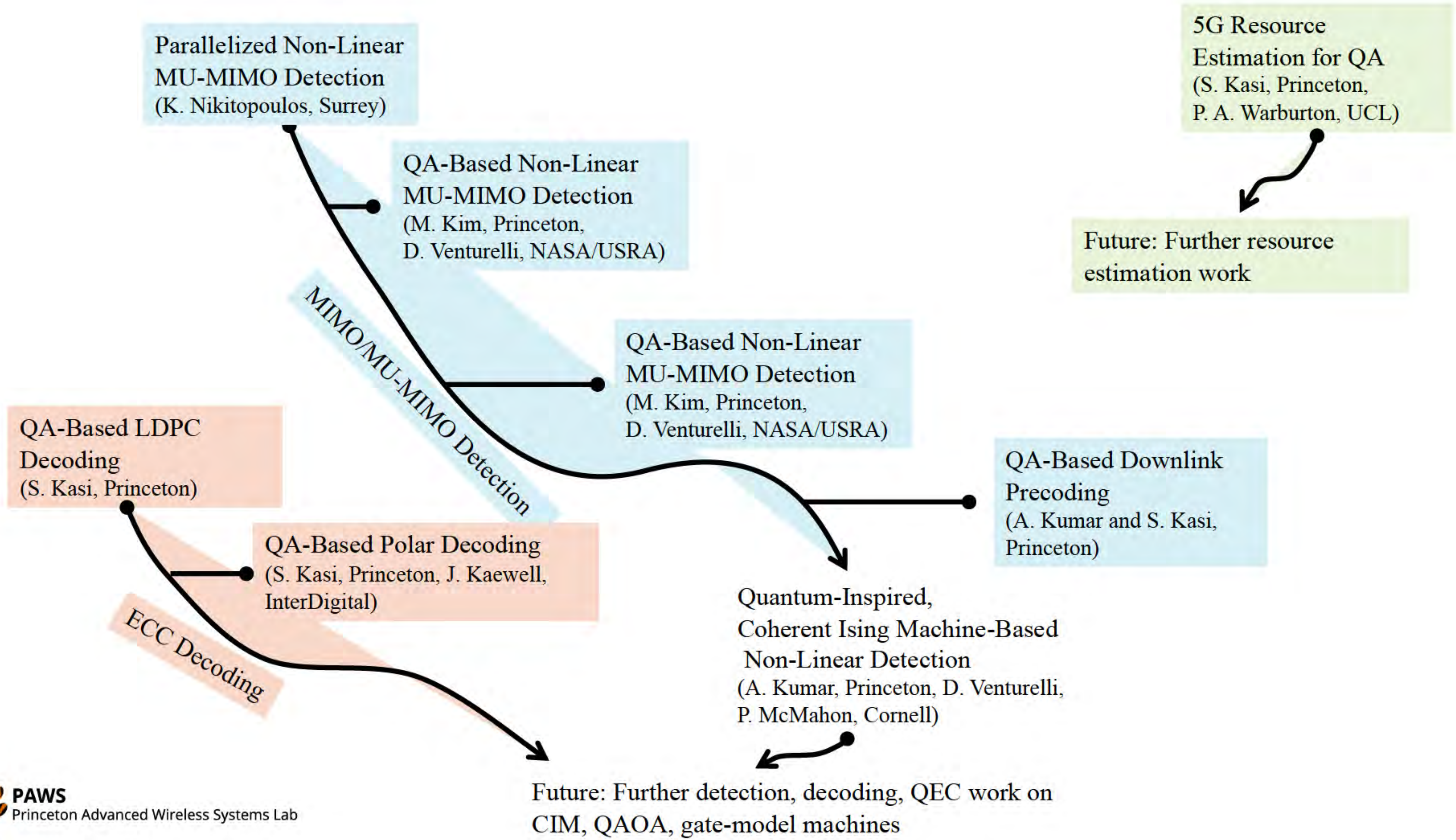
Why Quantum Compute for Wireless?

- Performance-compute elasticity
  - Spectral efficiency *v.* compute
- Detection: Zero-Forcing < MMSE < Sphere Decoder
- Decoding: quantization levels ↑, iteration counts ↑

The Physics Perspective

Why Wireless Applications?

- Must operate at "line rate"
- High computational throughput required
- Low computational latency required



Parallelized Non-Linear MU-MIMO Detection (K. Nikitopoulos, Surrey)
QA-Based Non-Linear MU-MIMO Detection (M. Kim, Princeton, D. Venturelli, NASA/USRA)
QA-Based Non-Linear MU-MIMO Detection (M. Kim, Princeton, D. Venturelli, NASA/USRA)
MIMO/MU-MIMO Detection
QA-Based Downlink Precoding (A. Kumar and S. Kasi, Princeton)
Quantum-Inspired, Coherent Ising Machine-Based Non-Linear Detection (A. Kumar, Princeton, D. Venturelli, P. McMahon, Cornell)
5G Resource Estimation for QA (S. Kasi, Princeton, P. A. Warburton, UCL)
Future: Further resource estimation work
QA-Based LDPC Decoding (S. Kasi, Princeton)
QA-Based Polar Decoding (S. Kasi, Princeton, J. Kaewell, InterDigital)
ECC Decoding
Future: Further detection, decoding, QEC work on CIM, QAOA, gate-model machines

Summary and Conclusion

1. Quantum LDPC decoder (*QBP*, MobiCom'20)
2. Energy-performance analysis (ISCA QRE, arXiv '22)
3. Uplink MU-MIMO detection (*IoT-ResQ*, MobiCom '22)

- Future work:
    - **New hardware:** Physics-Inspired H/W, Optical and Analog Machines, Hybrids
    - **New Problems:** different error control codes, further comms system parts

PAWS
Princeton Advanced Wireless Systems Lab

paws.princeton.edu